In the Claims:

Claims 1-34 are pending in this application, and the status of each is listed below.

1. (currently amended) A computer assisted method of auditing a superset of training data, the superset comprising examples of documents having one or more preexisting category assignments, the method including:

partitioning the superset into at least two disjoint sets, including a test set and a training set, wherein the test set includes one or more test documents and the training set includes examples of documents belonging ~~belong~~ to at least two categories;

automatically categorizing the test documents using the training set;

calculating a metric of confidence based on results of the categorizing step and comparing the automatic category assignments for the test documents to the preexisting category assignments; and

reporting the test documents and preexisting category assignments that are suspicious and the automatic category assignments that appear to be missing from the test documents, based on the metric of confidence.

2. (original) The method of claim 1, further including repeating the partitioning, categorizing and calculating steps until at least one-half of the documents in the superset have been assigned to the test set.

3. (original) The method of claim 2, wherein the test set created in the partition step has a single test document.

4. (original) The method of claim 2, wherein the test set created in the partition step has a plurality of test documents.

5. (original) The method of claim 1, further including repeating the partitioning, categorizing and calculating steps until substantially all of the documents in the superset have been assigned to the test set.

6. (original) The method of claim 1, wherein the partitioning, categorizing and calculating steps are carried out substantially without user intervention.

7. (original) The method of claim 5, wherein the partitioning, categorizing and calculating steps are carried out substantially without user intervention.

8. (original) The method of claim 1, wherein the partitioning, categorizing, calculating and reporting steps are carried out substantially without user intervention.

9. (original) The method of claim 5, wherein the partitioning, categorizing, calculating and reporting steps are carried out substantially without user intervention.

10. (original) The method of claim 1, wherein the categorizing step includes determining k nearest neighbors of the test documents and the calculating step is based on a k nearest neighbors categorization logic.

11. (currently amended) The method of claim 10, wherein the metric of confidence is an unweighted measure of distance between ~~the~~ a particular test document and the examples of documents belonging to various categories.

12. (currently amended) The method of claim 11, where the unweighted measure includes application of a relationship $\Omega_0(\mathbf{d}_t, T_m) = \sum_{\mathbf{d}\in\{K(\mathbf{d}_t)\cap T_m\}} s(\mathbf{d}_t, \mathbf{d})$,

wherein

Ω_0 is a function of the particular test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and

s is a metric of distance between the particular test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}$, the certain sample documents being among a set of k nearest neighbors of the particular test document having category assignments to the various categories T_m.

13. (currently amended) The method of claim 10, wherein the metric of confidence is a weighted measure of distance between ~~the~~ a particular test

document and the examples of documents belonging to various categories, the weighted measure taking into account the density of a neighborhood of the test document.

14. (currently amended) The method of claim 13, wherein ~~where~~ the weighted measure includes application of a relationship

$$\Omega_1(\mathbf{d}_t, T_m) = \frac{\sum_{\mathbf{d}_1 \in \{K(\mathbf{d}_t) \cap T_m\}} s(\mathbf{d}_t, \mathbf{d}_1)}{\sum_{\mathbf{d}_2 \in K(\mathbf{d}_t)} s(\mathbf{d}_t, \mathbf{d}_2)}, \text{wherein}$$

Ω_1 is a function of the test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and

s is a metric of distance between the test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, the certain sample documents $\mathbf{d}_1$ being among a set of k nearest neighbors of the test document having category assignments to the various categories T_m and the certain sample documents $\mathbf{d}_2$ being among a set of k nearest neighbors of the test document.

15. (currently amended) The method of claim 1, wherein the ~~identifying~~ reporting step further includes filtering the test documents based on the metric of confidence.

16. (currently amended) The method of claim 15, wherein ~~the~~ filtering ~~step~~ further includes color coding the identified test documents based on the metric of confidence.

17. (currently amended) The method of claim 15, wherein ~~the~~ filtering ~~step~~ further includes selecting for display the identified test documents based on the metric of confidence.

18. (original) The method of claim 1, wherein ~~the user interface is~~ reporting includes generating a printed report.

19. (original) The method of claim 1, wherein ~~the user interface is~~ reporting includes generating a file conforming to XML syntax.

20. (original) The method of claim 1, wherein ~~the user interface is~~ reporting includes generating a sorted display identifying at least a portion of the test documents.

21. (original) The method of claim 1, further including calculating a precision score for the identified test documents.

22. (currently amended) A computer assisted method of auditing a superset of training data, the superset comprising examples of documents having one or more preexisting category assignments, the method including:

determining k nearest neighbors of the documents in a test subset automatically partitioned from the superset;

automatically categorizing the documents based on the k nearest neighbors into a plurality of categories;

calculating a metric of confidence based on results of the categorizing step and comparing the automatic category assignments for the documents to the preexisting category assignments; and

reporting the documents in the test subset and preexisting category assignments that are suspicious and the automatic category assignments that appear to be missing from the documents in the test subset, based on the metric of confidence.

23. (original) The method of claim 22, wherein the metric of confidence is an unweighted measure of distance between the test document and the examples of documents belonging to various categories.

//

//

24. (original) The method of claim 23, where the unweighted measure includes application of a relationship $\Omega_0(\mathbf{d}_t, T_m) = \sum_{\mathbf{d} \in \{K(\mathbf{d}_t) \cap T_m\}} s(\mathbf{d}_t, \mathbf{d})$, wherein

> Ω_0 is a function of the test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and
>
> s is a metric of distance between the test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}$, the certain sample documents being among a set of k nearest neighbors of the test document having category assignments to the various categories T_m.

25. (original) The method of claim 22, wherein the metric of confidence is a weighted measure of distance between the test document and the examples of documents belonging to various categories, the weighted measure taking into account the density of a neighborhood of the test document.

26. (original) The method of claim 25, wherein the weighted measure includes application of a relationship $\Omega_1(\mathbf{d}_t, T_m) = \frac{\sum_{\mathbf{d}_1 \in \{K(\mathbf{d}_t) \cap T_m\}} s(\mathbf{d}_t, \mathbf{d}_1)}{\sum_{\mathbf{d}_2 \in K(\mathbf{d}_t)} s(\mathbf{d}_t, \mathbf{d}_2)}$, wherein

> Ω_1 is a function of the test document represented by the a feature vector $\mathbf{d}_t$ and of various categories T_m; and
>
> s is a metric of distance between the test document feature vector $\mathbf{d}_t$ and certain sample documents represented by feature vectors $\mathbf{d}_1$ and $\mathbf{d}_2$, the certain sample documents $\mathbf{d}_1$ being among a set of k nearest neighbors of the test document having category assignments to the various categories T_m and the certain sample documents $\mathbf{d}_2$ being among a set of k nearest neighbors of the test document.

27. (original) The method of claim 22, wherein the determining, categorizing and calculating steps are carried out substantially without user intervention.

28. (currently amended) The method of claim 22, wherein the ~~identifying~~ reporting step further includes filtering the documents based on the metric of confidence.

29. (original) The method of claim 28, wherein the filtering step further includes color coding the ~~identified~~ reported documents based on the metric of confidence.

30. (original) The method of claim 28, wherein the filtering step further includes selecting for display the ~~identified~~ reported documents based on the metric of confidence.

31. (original) The method of claim 22, wherein ~~the user interface is~~ reporting includes generating a printed report.

32. (original) The method of claim 22, wherein ~~the user interface is~~ reporting includes generating a file conforming to XML syntax.

33. (original) The method of claim 22, wherein ~~the user interface is~~ reporting includes generating a sorted display identifying at least a portion of the documents.

34. (original) The method of claim 22, further including calculating a precision score for the ~~identified~~ reported documents.